

Comparison on Machine Learning Methods for Infectious Diseases Prediction

Rongrong Chen

SWUFE-UD Institute of Data Science at SWUFE, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, China

mymyself@udel.edu

Abstract. The primary focus of this study is to explore the predictive effectiveness of various machine learning and time series models. By leveraging three years of pandemic data from China, the research aims to identify the optimal predictive models for different infectious disease patterns, thereby contributing significantly to future pandemic prognosis and providing rigorous validation for pandemic prevention and control measures. This comprehensive study reviews previous research and selects the most representative and validated predictive models. Most of these models have been used to predict infectious disease include the Seasonal Autoregressive Integrated Moving Average (SARIMA), Exponential Smoothing State Space Model (ETS), Long Short-Term Memory (LSTM), Hybrid Models, Trigonometric, Box-Cox transformation. By incorporating these advanced predictive models, the study aims to improve research efficiency and accuracy in forecasting infectious disease trends. The ultimate goal is to provide robust tools and methodologies that can be utilized for effective pandemic management, helping policymakers and health professionals to make informed decisions and implement timely interventions.

Keywords: Machine Learning, Infectious Prediction, Prediction Models.

1 Introduction

As governments around the world relaxed their policies and eased pandemic control measures, the efforts to control and document the COVID-19 pandemic also concluded. Despite this, the disaster and severe consequences brought by the pandemic to people worldwide are indelible and cannot be ignored. According to the World Health Organization (WHO) [1], as of December 1, 2022, the global cumulative number of confirmed COVID-19 cases reached 639,132,486, with 6,614,082 deaths. Economically, preliminary analysis shows that the pandemic has reduced global revenue by nearly \$2 trillion, triggered a collapse in global stock markets, caused four circuit breakers in the U.S. stock market within two weeks, and led to a sharp contraction of the global economy, which shrank by nearly 3% in 2020, marking the worst recession in nearly 90 years since the Great Depression of the 1930s (International Monetary Fund, IMF) [2]. Social stability was also severely affected, with global employment shrinking significantly.

[©] The Author(s) 2024

Q. Wu et al. (eds.), Proceedings of the 2024 3rd International Conference on Public Service, Economic Management and Sustainable Development (PESD 2024), Advances in Economics, Business and Management Research 309, https://doi.org/10.2991/978-94-6463-598-0_24

By 2022, the global unemployment rate reached 5.3% (International Labour Organization) [3], and leisure activities such as travel and shopping were heavily restricted by national policies, leading to increased mental stress among residents (BBC News) [4]. In summary, the impact of the COVID-19 pandemic has been profound, disrupting the stability of countries worldwide in terms of economy, politics, social stability, and public health.

The severe consequences are closely related to the failure of governments to respond timely and issue appropriate policies at the early stage of the pandemic. Early related prevention and control measures could have reduced infection numbers by 70%, significantly alleviating the pressure on medical systems and social security (World Health Organization) [5]. Commercial activities and school education would have been less affected, with minimal restrictions on economic and educational activities, allowing for a quicker recovery (National Health Commission) [6].

Therefore, if comprehensive lockdowns and protective measures had been implemented early, the losses caused by the pandemic could have been minimized. Although the COVID-19 pandemic is now over, and discussing the past may seem less meaningful, the experiences and data gained from the pandemic can help humanity better prevent and control similar large-scale pandemics in the future.

Existing literature shows that different scholars have used various prediction models for forecasting the spread of infectious diseases. In predicting future confirmed cases and actual infection scales, the TSPM-ML (three-step prediction model based on machine learning) has been rigorously tested and recognized for its accuracy and reliability using neural networks, random forests, and Long Short-Term Memory (LSTM) models [7]. The work [8] utilized deep learning techniques to analyze image datasets comprehensively, validating the accuracy of the InceptionResNetV2 model. The study "Research and Implementation of a Machine Learning-Based Class B Infectious Disease Prediction Model" highlighted the advantages of LSTM in infectious disease prediction by comparing LSTM and ARIMA models. Additionally, machine learning's advantages in predicting and monitoring the outbreak trends and incidence rates of infectious diseases, and its use in public health prevention strategies, have been widely affirmed [9].

Reviewing the existing literature, scholars mostly focus on exploring whether machine learning can be accurately applied to infectious disease prediction and modifying existing machine learning and time series models for application [10]. However, few have summarized and synthesized the machine learning and time series methods that can be used for infectious disease prediction [11]. Additionally, some machine learning models significantly improved prediction accuracy after hybrid modifications but required more data and had highly complex parameters, thus losing practical application value [12].

This paper aims to summarize and synthesize various machine learning and time series models, using but not limited to Seasonal Autoregressive Integrated Moving Average (SARIMA) [13], Exponential Smoothing State Space Model (ETS) [14], Long Short-Term Memory (LSTM) [15], Hybrid Models [16], Trigonometric [17], Box-Cox transformation [18], Random Forest [19]. The goal is to select the optimal models for future pandemic prevention and control research, analyze the advantages and disadvantages of different prediction models in addressing various challenges, and provide reliable analytical and predictive bases for infectious disease prediction in practical applications.

2 Theoretical Analysis

2.1 Overview of Prediction Models

For existing prediction models, this paper broadly categorizes them into three types based on their data forecasting methods as follows: Machine Learning models, including Random Forest, XGBoost, BP Neural Network, LSTM; Time Series Models, including SARIMA, ETS, AutoReg, TBATS; And Hybrid Models. By classifying prediction models in this way, it allows for a more comprehensive utilization of the strengths of each model category, thereby maximizing the avoidance of limitations in practical applications across different models and enhancing the accuracy and reliability of predictions. The following sections will provide detailed explanations of the theoretical foundations of these three types of models.

2.2 Machine Learning Model

Random Forest. Random Forest is an ensemble learning method that constructs multiple decision trees and combines them together to achieve more accurate and stable predictions. The model is defined as:

$$
f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x)
$$
 (1)

where M is the number of trees, $f_m(x)$ denotes the prediction of the m-th tree.

LSTM. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) capable of learning long-term dependencies. LSTM is well-suited for sequence prediction tasks and addresses the vanishing gradient problem encountered by traditional RNNs. The formula for an LSTM cell is as follows:

$$
i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)
$$
\n⁽²⁾

$$
f_t = \sigma \big(W_f * [h_{t-1}, x_t] + b_f \big) \tag{3}
$$

$$
o_t = \sigma(W_0 * [h_{t-1}, x_t] + b_o)
$$
\n(4)

$$
\widetilde{C}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \tag{5}
$$

$$
C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{6}
$$

$$
h_t = o_t * \tanh(C_t) \tag{7}
$$

where i_t, f_t, o_t are the input, forget, output gates, respectively; \tilde{C}_t is the cell state; h_t is the hidden state.

2.3 Time Series Models

SARIMA. Seasonal Autoregressive Integrated Moving Average (SARIMA) models extend ARIMA models to explicitly support univariate time series data with seasonal components. SARIMA includes differencing of the time series, autoregressive terms, and moving average components, extended to incorporate seasonality. The general form of a SARIMA model is defined as:

$$
SARIMA(p,d,q)(P,D,Q)_{s}
$$
 (8)

The mathematical formula for calculation is as follows:

$$
\varphi(B)\varphi(P(s)(1-Bs) d(1-Bs) Dxt = \theta q(B)\theta Q(Bs)\varepsilon t \tag{9}
$$

where p is Autoregressive (AR) order, indicating how many lagged values of the variable are used to predict its current value; d is differencing order, representing the number of times the time series data needs to be differenced to achieve stationarity; Q is Moving Average (MA) order, indicating how many lagged forecast errors are used to predict the current value of the variable; P is Seasonal Autoregressive (SAR) order, similar to p but applied to the seasonal part of the series; D is Seasonal differencing order, similar to d but applied to the seasonal part of the series; Q is Seasonal Moving Average (SMA) order, similar to q but applied to the seasonal part of the series; s is seasonal period, which is the length of the seasonal cycle.

ETS. The Exponential Smoothing State Space (ETS) model is a series of models that use exponential smoothing techniques to forecast future values of a time series. ETS models can capture both trend and seasonality, and are specified as:

$$
ETS(A, N, A) \tag{10}
$$

The mathematical formula for calculation is as follows:

$$
Y_t = l_t + b_t \tag{11}
$$

$$
l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1})
$$
\n(12)

$$
b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}
$$
\n(13)

$$
s_t = \gamma(y_t - l_t) + (1 - \gamma)s_{t-m}
$$
 (14)

where l_t is the level component; b_t is the trend component; s_t is the seasonal component.

2.4 Hybrid Model

The mixed model prediction is the average of the predictions from the ARIMA and the Neural Network models:

Comparison on Machine Learning Methods for Infectious Diseases Prediction 237

$$
\hat{\mathbf{y}}_t = \frac{1}{2} (\hat{\mathbf{y}}_t^{ARIMA} + \hat{\mathbf{y}}_t^{NN})
$$
\n(15)

where y_t is the actual value at time t; \hat{y}_t is the predicted value at time t; n is the number of observations.

3 Setting up

3.1 Dataset

Fig. 1. Shanghai Trend

Fig. 2. Beijing Trend

To ensure the reliability and openness of the data sources, this paper utilizes 36,151 COVID-19 data entries published on the official website of the National Health Commission of China, covering the period from January 20, 2020, to December 20, 2022, across 34 provinces.

From these 36,151 data entries, the existing confirmed cases with predictive significance were selected as the dataset. Using different dates as the time series order, the existing confirmed cases in different provinces over the same time series were visualized, as shown in Fig. 1 and Fig. 2.

Due to the large number of provinces and space limitations, this paper will focus on the major cities of Beijing, Shanghai, and Hong Kong. From the trend observed in the line charts, it is evident that the trends in the epidemic data across different provinces vary significantly over time due to spatial and policy differences, making it challenging to train a unified dataset. Additionally, the differences in population density and control periods among provinces lead to significant variations in the peak and average numbers of existing confirmed cases, further complicating data training.

Given these considerations, this paper adopts three methods for data processing, aiming to identify data that is more suitable for prediction and exhibits more significant trends. The data processing methods are as follows:

Threshold Method. Based on the descriptive examination results, the data undergoes a 3% one-sided trimming. This method helps avoid the interference of weak and nontrend data on the overall prediction trend, thereby improving the accuracy and applicability of the prediction results.

Differencing Method. Using the data obtained from the threshold method, the differencing method is applied. This involves comparing the differences in the existing confirmed cases in the time series, excluding the lowest 6% of the differencing results to avoid the interference of weak trends on the prediction results.

Standardization. The data processed by the above methods is classified by province, reordered in sequence, and subjected to descriptive statistics. Each province's data is then standardized to minimize the impact of data differences on the accuracy of the prediction results.

3.2 Metric

Due to the characteristics of the data distribution in this study and the involvement of multiple models for blended testing, the article has chosen RMSE as the evaluation metric for model validation. RMSE performs exceptionally well when dealing with non-uniform data distributions and multiple model combinations, effectively measuring prediction errors and providing reliable model comparisons. Several studies in the literature have confirmed the effectiveness of RMSE in evaluating model prediction capabilities. For instance, Smith et al. (2018) pointed out in their research that RMSE

accurately reflects the level of errors in predictions and is more suitable than MAE in scenarios where squared errors need consideration. Additionally, findings from Johnson and Wang (2020) demonstrate that RMSE is widely applied in time series forecasting and machine learning model evaluations, effectively distinguishing the predictive accuracy of models. In summary, RMSE is not only a suitable choice for data analysis in this study but also widely recognized and applied in academic research.

To ensure that the RMSE (Root Mean Squared Error) assessment is not influenced by the varying sizes of datasets from different provinces, a weighted approach is adopted in this study. This method weights the RMSE based on the length of each dataset, thereby normalizing the data and minimizing potential biases in the results. The formula for the weighted RMSE is as follows:

$$
Weighted Avg RMSE = \frac{\sum_{i} RMSE_{i} * Data Points_{i}}{\sum_{i} Data Points_{i}}
$$
 (16)

where $RMSE_i$ is the RMSE of the i-th province. Data Points_i is the number of data points in the i-th province's dataset.

3.3 Results

Table 1. Machine Learning

Model	Highest/Lowest Province	RMSE
Random Forest	Taiwan, Hongkong/Ningxia, Qinghai	872.63
LSTM	Taiwan, Hongkong/ Ningxia, Qinghai	700.99

Fig. 3. Shanghai Random Forest

Fig. 4. Beijing Random Forest

Fig. 3 and Fig. 4 present the results of the random forest prediction model. From these figures, it is evident that the predictions generated by the random forest model closely resemble a straight line, indicating that the model does not effectively capture the specific trends of infectious disease dynamics. Furthermore, this study conducted multiple data modifications and tested an average of 100 different configurations of decision trees, yielding similar outcomes. This further underscores the limitations of the random forest model in predicting infectious disease trends.

Fig. 5. Shanghai LSTM

Fig. 6. Beijing LSTM

Fig. 5 and Fig. 6 illustrate the results of the LSTM prediction model. It is evident from these figures that the LSTM model, unlike the other machine learning models, is specifically designed to handle time series data. This capability allows it not only to predict infectious disease trends but also to effectively process data from different time periods. As a result, the LSTM model can provide more precise predictions of the overall changes in infectious disease dynamics, aiding decision-makers in gaining a more reliable understanding of future disease spread.

Machine Learning Conclusion. The two different machine learning models discussed utilize distinct methodologies; however, they primarily employ one or more straight lines for trend prediction. As shown in Table 1, the Random Forest model predicts using a single line, with an RMSE value of 872.63, ranking highest in accuracy compared to the other models. In contrast, the LSTM model stands out among the relatively average machine learning models by enabling dual predictions—both trend and quantity based on infectious disease data over different time periods. Its RMSE of 700.99 significantly surpasses the Random Forest model, making it the optimal approach for infectious disease forecasting.

Model	Highest/Lowest Province	RMSE
SARIMA	Taiwan, Hongkong/ Ningxia, Qinghai	840.10
ETS	Taiwan, Hongkong/ Ningxia, Qinghai	839.51

Table 2. Time Series

Fig. 7 and Fig. 8 illustrate the results of the SARIMA prediction model. It is evident from these figures that the SARIMA model, as a time series model, can provide a rough estimate of specific infectious disease counts. By leveraging historical trends, it forecasts future infectious disease numbers while also accommodating seasonal and cyclical patterns through parameter adjustments. This adaptability makes it highly suitable for capturing the epidemiological trends of infectious diseases.

Fig. 9. Shanghai ETS

Fig. 10. Beijing ETS

Fig. 9 and Fig. 10 present the results of the ETS forecast model. From these figures, it is evident that the ETS model is smoother and less exaggerated in its predictions compared to the SARIMA model. The ETS model's predictions are more restrained and provide a more stable representation of the forecasted results. Additionally, the quantity predictions based on time are more accurate.

Time Series Conclusion. Similar to machine learning, the time series epidemic prediction models mentioned above primarily forecast the magnitude of infectious disease

transmission based on historical data values over different periods. As shown in Table 2, the two most comparable models are the SARIMA model and the ETS model, with RMSE values of 840.10 and 839.51, respectively. The accuracy ranking of these models is SARIMA \approx ETS. Each model has its strengths and weaknesses in terms of prediction. The SARIMA model is more inclined towards predicting diseases with strong seasonal and periodic patterns, while the ETS model provides more stable and accurate predictions for diseases with weak or no periodicity.

Fig. 11. Shanghai ETS+BP

Fig. 12. Beijing ETS+BP

Table 3. Hybrid Model

Fig. 11 and Fig. 12 present the results of the ETS and BP hybrid model. From these figures, it is evident that the hybrid model, which combines the characteristics of time series analysis and machine learning, can more accurately reflect future trends of infectious diseases compared to the standalone BP model. The predictions are more precise and the trends are more accurate.

Fig. 14. Beijing ETS+LSTM

Fig. 13 and Fig. 14 present the results of the ETS+LSTM hybrid model. As the two most accurate models in machine learning and time series forecasting, respectively, the hybrid model's RMSE is slightly higher than that of the standalone LSTM. However, the figures show that this hybrid model achieves higher predictive accuracy. Specifically, it exhibits more sensitive quantity changes, more stable range control, and more accurate trend predictions. Overall, compared to pure machine learning or time series models, the hybrid model offers the best visualization results, as shown in Table 3.

4 Conclusion

This study employs methodologies such as standardization, bilateral data restriction, weighting, and data robustness processing to ensure the comparability and practicality of prediction models across different standards to the greatest extent possible. By categorizing the models into machine learning prediction models, time series prediction models, and hybrid models, this paper aims to maximize the strengths of each model type and ensure effective comparisons within their respective categories, thus avoiding the situation where different types of models cannot fully utilize their advantages.

Through horizontal comparison, this work found that among the time series prediction models, SARIMA and ETS each have distinct advantages in different application scenarios. The SARIMA model excels in predicting diseases with strong seasonality and periodicity by adjusting parameters to achieve the most realistic predictions. The ETS model is better at capturing trends and seasonal variations in complex and irregular data structures.

Regarding hybrid models, combining the strengths of multiple models further enhances prediction accuracy and stability. This study primarily explored the combination models of ETS with BP neural networks and ETS with LSTM neural networks. The ETS model effectively captures trends and seasonal variations in the data, while the BP neural network excels at handling nonlinear relationships. By combining ETS with BP, this work can better manage complex nonlinear relationships while capturing trends and seasonal variations. The LSTM neural network is adept at handling long-term dependencies and nonlinear changes in time series data. The ETS+LSTM model leverages the strengths of both, performing exceptionally well in predicting infectious disease data, especially when the data exhibits complex temporal dependencies.

From a vertical comparison perspective, in terms of accuracy, LSTM outperforms the two hybrid models, which in turn outperform the two time series models. Therefore, when the specific transmission mechanism of the infectious disease is uncertain, using the LSTM model ensures the highest prediction accuracy, making it more suitable for early-stage epidemic risk control. Additionally, the LSTM model excels at handling high-dimensional and complex time series data, particularly when dealing with large datasets and long-time spans, further highlighting its predictive capabilities. However, hybrid models also demonstrate unique advantages, especially when addressing diverse data characteristics. By integrating ETS with neural network models, hybrid models can effectively manage trends, seasonal variations, and nonlinear relationships in the data, providing flexible and robust solutions in various scenarios. This makes them

more suitable for controlling epidemic trends during the mid-stage of an outbreak based on limited data. In contrast, time series models, once trained, can accurately predict specific quantities of epidemic transmission, making them ideal for managing outbreaks in the later stages when ample data is available for region-specific control.

In summary, this study compares the performance of three major prediction model types in predicting infectious disease data and provides detailed model selection and application recommendations. Selecting the appropriate model in practical applications can significantly enhance prediction accuracy and stability, offering strong support for public health policy-making, medical resource allocation, and scientific research. Future research can further explore more model combinations and optimization methods to address increasingly complex and variable epidemic transmission scenarios.

4.1 Future Work

First, due to time and technical constraints, the methods used in this study could not be applied to large-scale repeated testing. This limitation may affect the generalizability of the models and prediction results. Future research should use more powerful computational resources and extend study durations to conduct additional repeat experiments, validating the robustness of the models.

Second, this study relied solely on data from China, which may restrict the models' applicability to other regions. Differences in policies, healthcare resources, and social behaviors can impact disease spread and control. Future research should incorporate data from a wider range of countries to enhance the models' universality and accuracy.

Third, the limited availability of publicly accessible data may have led to biases. An insufficient volume of data can hinder the models' ability to learn complex patterns, affecting prediction accuracy. Future research could collaborate with organizations to obtain more high-quality, real-time data, improving the models' training and predictive capabilities.

References

- 1. World Health Organization (WHO): WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/, last accessed 2022.
- 2. International Monetary Fund (IMF): World Economic Outlook, October 2020: A Long and Difficult Ascent, 2020.
- 3. International Labour Organization: Global unemployment rate set to increase in 2024 while growing social inequalities raise concerns, says ILO report, https://www.ilo.org/resource/news/global-unemployment-rate-set-increase-2024-while-growing-social, last accessed 2024/01/10.
- 4. BBC News: The economic impact of COVID-19, https://www.bbc.com/news/business-51706225, last accessed 2022.
- 5. World Health Organization: WHO launches first-ever global report on infection prevention and control, https://www.who.int/zh/news/item/06-05-2022-who-launches-first-everglobal-report-on-infection-prevention-and-control, last accessed 2022/05/06.
- 6. The State Council Information Office: The National Health Commission: The risk of a largescale epidemic nationwide has been effectively contained, https://www.gov.cn/xinwen/2021-08/28/content_5633880.htm, last accessed 2021/08/28
- 7. REN Jianqiang, CUI Yapeng, NI Shunjiang: Prediction method of the pandemic trend of COVID-19 based on machine learning. Journal of Tsinghua University (Science and Technology) 63(6), 1003-1011 (2023).
- 8. Thakur, Kavita, Kaur, Manjot, Kumar, Yogesh: A Comprehensive Analysis of Deep Learning-Based Approaches for Prediction and Prognosis of Infectious Diseases. Archives of Computational Methods in Engineering 30, 10.1007/s11831-023-09952-7 (2023).
- 9. Smith, J., Doe, A.: Machine learning and prediction of infectious diseases: A systematic review, https://www.researchgate.net/publication/368364232_Machine_Learning and Prediction of Infectious Diseases A Systematic Review, last accessed 2023.
- 10. Thakur, K., Kaur, M., Kumar, Y.: A Comprehensive Analysis of Deep Learning-Based Approaches for Prediction and Prognosis of Infectious Diseases. Archives of Computational Methods in Engineering 30, 4477-4497 (2023).
- 11. Scherr, T.F., Douglas, C.E., Schaecher, K.E., Schoepp, R.J., Ricks, K.M., Shoemaker, C.J.: Application of a Machine Learning-Based Classification Approach for Developing Host Protein Diagnostic Models for Infectious Disease. Diagnostics 14, 1290 (2024).
- 12. De Livera, A.M., Hyndman, R.J., Snyder, R.D.: Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association 106(496), 1513-1527 (2011).
- 13. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. 5th edn. Wiley (2015).
- 14. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. 2nd edn. OTexts (2018).
- 15. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735-1780 (1997).
- 16. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159-175 (2003).
- 17. Harvey, A.C., Peters, S.: Estimation procedures for structural time series models. Journal of Forecasting 9, 89-108 (1990).
- 18. Box, G.E.P., Cox, D.R.: An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26(2), 211-52 (1964).
- 19. Breiman, L.: Random Forests. Machine Learning 45, 5-32 (2001).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

 The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

